ANALYZING MULTIPLE-STEP REASONING ABILITIES OF TRANSFORMERS

Anonymous authors
Paper under double-blind review

CONTENTS

A VISUALIZATION OF TRANSFORMER LAYERS

Following Piotrowski et al. (2019), we also attempt to understand what the Transformer networks are learning through layer-wise visualization of attention (Vig, 2019). We take model trained on COARSE granularity proofs using INFIX representation for 1 variable using the SMALL COEFF configuration. We take the following example:

$$\begin{aligned} P_0 &= \underline{(4 * x_1^2) * (5 * x_1^3 + 4 * x1)} + (12 * x_1), && \text{/* MULSTEP */} \\ &= (20 * x_1^5 + 16 * x_1 * *3) + (12 * x_1) \end{aligned}$$

In Figure 1, we observe that in layer 2 encoder-decoder attention indicates that while generating the number 16, the Transformer network is clearly able to attend to the two digits 4 and 4 required for the multiplication. In Figure 2, we observe that the Transformer networks, in the same time also learns to copy the expression $12 + x_1$ in Layer 1. Even though such clear logical patterns emerge quite frequently, in some cases patterns become hard to interpret.

B ALGORITHMS

The polynomial sampling algorithms `buildProduct` and `buildFactor` are provided in Algorithms 1 and 2 respectively.

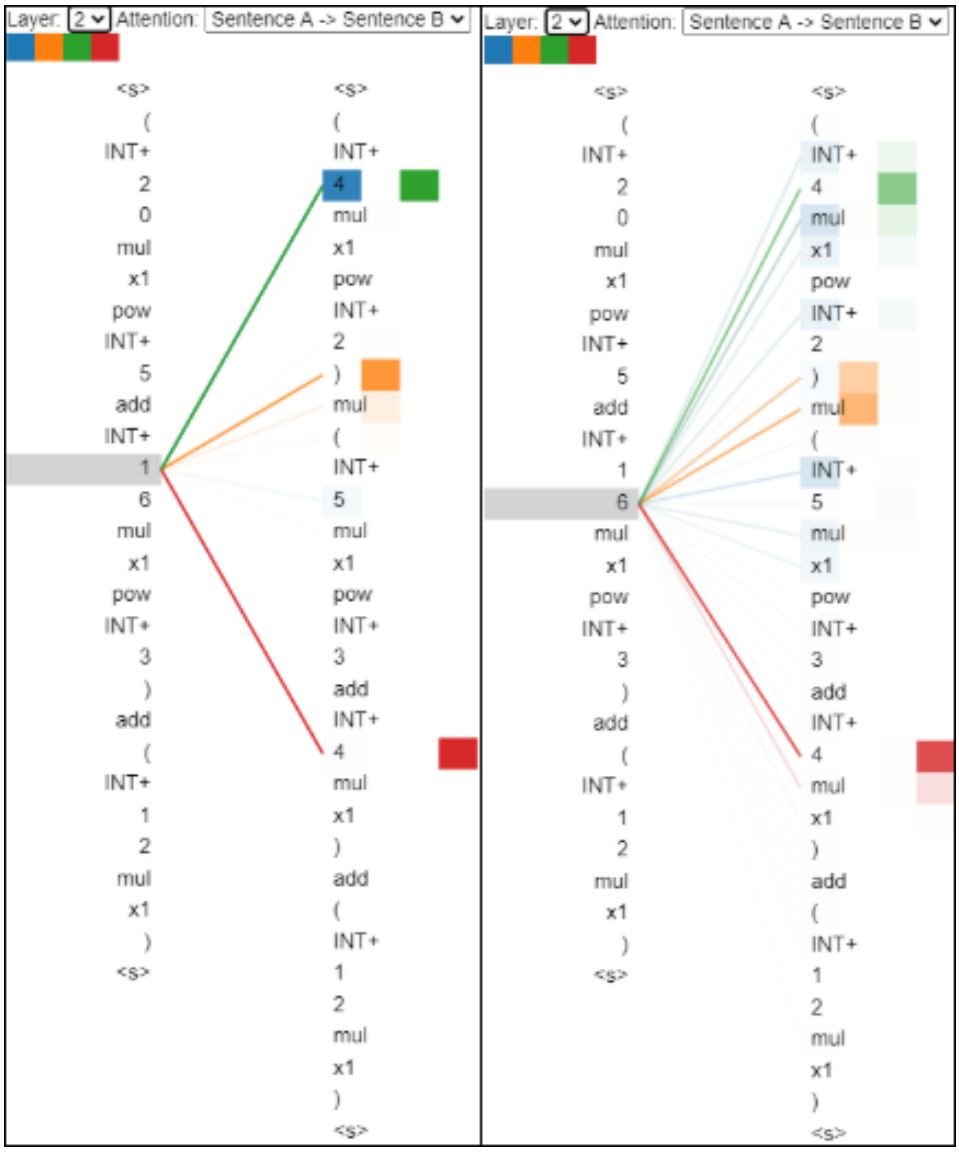

Figure 1: The layer 2 encoder-decoder attention for the output digits 16 in the first simplified product for the output $(20 * x_1^5 + 16 * x_1 * *3) + (12 * x_1)$. As expected, the digits 1 and 6 attends to the coefficients of the first and third monomial in the input expression $\underline{(4 * x_1^2) * (5 * x_1^3 + 4 * x1)} + (12 * x_1)$. Config: COARSE, SMALL COEFF, INFIX, 1 variable.

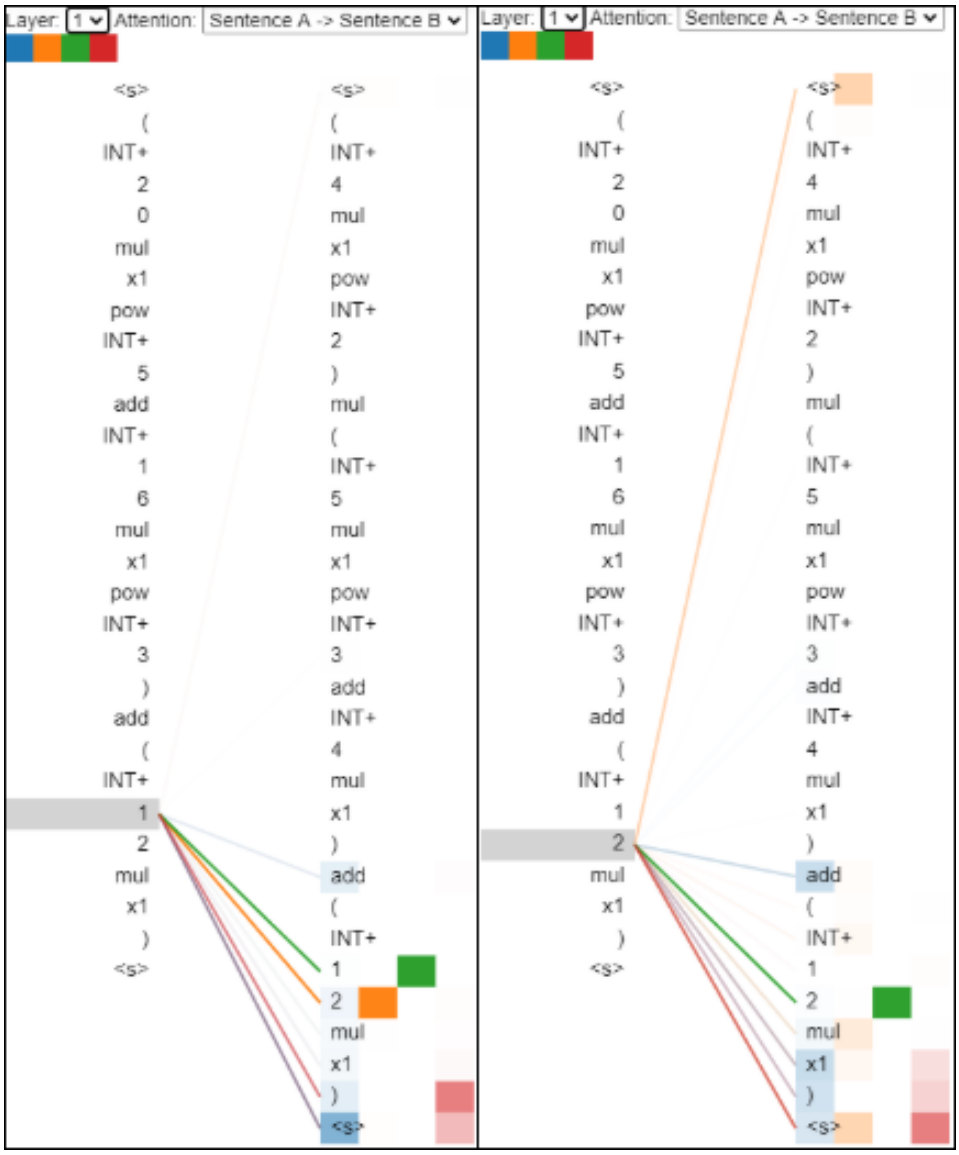

Figure 2: The layer 1 encoder-decoder attention for the coefficient 12 in the last product $(20 * x_1^5 + 16 * x_1 * *3) + \underline{(12 * x_1)}$. It is expected, that in this step, this product remains unchanged and simply copied o the output. Therefore, we see that the layers learn to copy the coefficients directly by attending to the corresponding digits (i.e. 1 attends to 1 in the last product). Config: COARSE, SMALL COEFF, INFIX, 1 variable.

Algorithm 1: BuildProduct (Sampling Products)

Input: $x_{\mathcal{P}}$, mdeg
Constraints: nvars_prod, max_coeff_prod, max_fac_prod, max_terms_prod
Output: A list of factors F_{seq}

Sample $nvar \in \{\texttt{num\_vars\_fac}, \ldots, \texttt{nvars\_prod}\}$
$nvar = min(|x_{\mathcal{P}}|, nvar)$
Sample $nvar$ variables from $x_{\mathcal{P}}$ as $x_{\mathcal{P}_i}$ // Variable set for this product
Sample $nfac \in \{2, \ldots, \texttt{max\_fac\_prod}\}$ // #Factors for this product
/* Get maximum degree, terms and coefficient available */
$rdegree$ = $mdeg$, $rterms$ = max_terms_prod, $rcoeff$ = max_coeff_prod
$cprod$ = 1 // Keeping track of product built till now
$F_{seq} = [\,]$
for $j \leftarrow 1$ **to** $nfac$ 1 **do**
  $f_j = \texttt{buildFactor}(x_{\mathcal{P}_i}, rdegree, rterms, rcoeff)$
  /* Update degree, terms and coefficient for next factor */
  $cprod = cprod * f_j$
  $rdegree = rdegree - degree(f_j)$
  $rterms$ = max_terms_prod/$|terms(cprod)|$
  $rcoeff$ = max_coeff_prod/$max(coeffs(cprod))$
  Append f_j in F_{seq}
  if $rdegree == 0$ **then**
    break
  end
end
Shuffle F_{seq}

Algorithm 2: BuildFactor (Sampling A Factor)

Input: $x_{\mathcal{P}_i}, rdegree, rterms, rcoeff$
Constraints: num_vars_fac, max_coeff_fac, max_terms_fac, max_degree_fac
Output: A factor f_j, Number of terms $nterms_j$

Sample $nvar \in \{1, \ldots, \texttt{num\_vars\_fac}\}$
$cvars$ = Sample $nvar$ variables from $x_{\mathcal{P}_i}$ // Variable set for this factor
Sample $nterms \in \{1, \ldots, min(\texttt{max\_terms\_fac}, rterms)\}$ // # Terms for this factor
Sample $\{d_k\}_{k=1}^{nterms}$, s.t. $d_k \in \{0, \ldots, min(\texttt{max\_degree\_fac}, rdegree)\}$ // Term degrees: degree 0 allows for constant terms
Sample $\{c_k\}_{k=1}^{nterms}$, s.t. $c_k \in \{1, \ldots, min(\texttt{max\_coeff\_fac}, rcoeff)\}$ // Term coefficients
for $k \leftarrow 1$ **to** $nterms$ 1 **do**
  selects $d[k]$ variables from $cvars$ with replacement
  // E.g. if $d[k] = 4, cvars = [x_1, x_2]$. May sample $[x_1, x_2, x_1, x_1]$
  Convert the selected $d[k]$ variables to a term // $t_k = c_k * x_1^3 * x_2$,
end
$f_j = \sum_{k=1}^{nterms} t_k$
return f_j;

C TABLE OF CONSTRAINTS AND NOTATIONS

We provide the full list of constraints and notations in Table 1.

Term Constraints	#Products	$\text{nprod} \in \{2, \ldots, \text{maxP}_{\mathcal{P}}\}$
	#Factors in P_i	$\text{nfac}_i \in \{2, \ldots, \text{maxf}_{\text{P}}\}, \forall i \in \{1, \ldots, \text{nprod}\}$
	#Terms in f_{ij}	$\lvert\text{terms}(f_i)\rvert \in \{1, \ldots, \text{maxT}_{\text{f}}\}, \forall f_i j \in f_{\mathcal{P}}$
	#Terms in $\hat{P}_i$	$\lvert\text{terms}(\hat{P}_i)\rvert \leq \text{maxT}_{\text{P}} \forall P_i \in \mathcal{P}_0$
Degree Constraints	#Degree in $\hat{\mathcal{P}}$	$\sum d_{mn} \leq \text{D}_{\mathcal{P}},\ \forall m\ \hat{t_m} \in \text{terms}(\hat{P}), \forall n\ x_n \in \text{vars}(\hat{t_m})$
	#Degree in f_{ij}	$\sum d_{kl} \leq \text{D}_{\text{f}}, \forall k\ \text{terms}(f_{ij}), \forall f_{ij} \in f_{\mathcal{P}}$
Variable Constraints	#Variables in $\mathcal{P}_0$	$\lvert x_{\mathcal{P}}\rvert \leq \text{V}_{\mathcal{P}}$
	#Variables in P_i	$\lvert\text{vars}(\text{P}_\text{i})\rvert \leq \text{V}_{\text{P}}, \forall P_i \in \mathcal{P}_0$
	#Variables in f_i	$\lvert\text{vars}(\text{f}_\text{ij})\rvert \leq \text{V}_{\text{f}}, \forall f_j \in f_{\mathcal{P}}$
Coefficient Constraints	Coeff in $\hat{\mathcal{P}}$	$\hat{a_j} \leq \text{C}_{\mathcal{P}}, \forall \hat{a_j} \in \text{coeffs}(\hat{P})$
	Coeff in $\hat{P}_i$	$\hat{a_{ij}} \leq \text{C}_{\text{P}}, \forall a\ \text{coeffs}(\hat{P}_i), \forall P_i \in \mathcal{P}_0$
	Coeff in f_i	$a_k \leq \text{C}_{\text{f}}, \forall a\ \text{coeffs}(f_{ij}), \forall f_{ij} \in f_{\mathcal{P}}$

Table 1: List of notations, and corresponding constraints that a generated polynomial satisfies.

D PROBLEM SPACE SIZE ESTIMATION

We present the problem space size estimates here in Table 2.

Config	NVAR = 1		NVAR = 2	
	Equation Size Estimate	Endpoint Size Estimate	Equation Size Estimate	Endpoint Size Estimate
SMALL COEFF	104M	8.24M	184M	27.4M
MEDIUM COEFF	179M	16.3M	325M	42.4M
LARGE COEFF	289M	32M	507M	68.8M
NO BACKTRACK	324M	54.9M	538M	104M
MEDIUM DEG	459M	67.4M	902M	144M
MEDIUM TERMS	866M	31.5M	1.73B	801M

Table 2: Size Estimates for the problem space, after generating sets of size 5M.

E INPUT REPRESENTATION (ADDITIONAL RESULTS)

We present the results for FINE configuration for 2 variable setting here in Table 3. The errors made by the models for 1 Variable and 2 Variable settings are presented in Tables 4 and 5 respectively.

F ANNOTATED PROOF (ADDITIONAL RESULTS)

We present the results for COARSE and FINE configuration for 2 variable setting for annotated proofs here in Table 6. The errors made by the models for 1 Variable and 2 Variable settings are presented in Tables 7 and 8 respectively.

G FULLY SYMBOLIC PROOFS

As $>$ 80% of the errors occurred in multiplication step, we separately tested the Transformer's ability to do arithmetic, by creating datasets involving multiplication and addition of 4-digit and 9-digit numbers. While the models quickly achieved an accuracy of close to 99% for addition; for multiplication, they could not go beyond even 1% after seeing 2M examples. Hence, we envision

Config	Proof Type	Endpoint		#Train	Full Proof		Stepwise		Calibration			
		#EE	Endpoint Acc.		Full Proof Acc.	Greedy Stepwise Acc.	Top-1 Acc.	Beam-5 Acc.	Sure Rate	P	R	F1
SMALL COEFF	Infix/Fine	4.3M	94.7	4.6M	88.1	97.19	90.7	92.2	83.47	100	92.02	0.96
	Prefix/Fine	4.5M	93.93	5.4M	90.3	97.83	94.63	96.2	87.9	99.96	92.85	0.96
MEDIUM COEFF	Infix/Fine	7M	95.3	4.4M	82.2	96.25	94.28	95.76	86.24	100	91.47	0.96
	Prefix/Fine	5.2M	92.77	2.9M	72.4	93.6	91.53	94.33	81.97	100	89.55	0.94
LARGE COEFF	Infix/Fine	9M	91.8	3.2M	73	93.85	77.94	82.2	63	99.9	80.75	0.89
	Prefix/Fine	6.1M	86.6	4.7M	78.6	95.6	91.93	93.47	83.87	100	91.23	0.95
NO BACKTRACK	Infix/Fine	8.6M	83.8	5.8M	72.5	94.64	81.54	84.82	72.34	100	88.72	0.94
	Prefix/Fine	7.1M	79.2	4.1M	60.7	90.48	81.73	85.67	70.2	99.91	85.81	0.92
MEDIUM DEG	Infix/Fine	4.9M	87.9	3.6M	73.5	94.21	89.78	92.46	77.22	100	86.01	0.92
	Prefix/Fine	5.2M	83.73	4.6M	73.6	94.57	86.5	89.4	76.93	100	88.94	0.94
MEDIUM TERMS	Infix/Fine	8.5M	90	4.8M	64	92.98	79.04	81.86	66.92	99.88	84.56	0.92
	Prefix/Fine	6.6M	87.07	4.5M	62.9	92.74	86.4	89.07	73.67	100	85.26	0.92

Table 3: Results for FINE configuration for 2 Variables for Infix and Prefix representation (No curriculum, No annotation).

Config	Proof Type	Full Proof		Error Percentage					
		Full Proof Accuracy	Greedy Stepwise Accuracy	First FacStep	Total FacStep	First MulStep	Total MulStep	First SumStep	Total SumStep
SMALL COEFF	Coarse/Infix	95	98.83	8	9.43	88	84.91	4	5.66
	Fine/Infix	**98.9**	**99.79**	0	0	100	100	0	0
	Coarse/Prefix	95.3	98.97	4.26	4.08	72.34	71.43	23.4	24.49
	Fine/Prefix	96.9	99.4	9.68	9.68	77.42	77.42	12.9	12.9
MEDIUM COEFF	Coarse/Infix	92.8	98.24	1.39	1.25	95.83	92.5	2.78	6.25
	Fine/Infix	90.3	97.99	11.34	11.32	85.57	84.91	3.09	3.77
	Coarse/Prefix	**93.6**	**98.58**	3.12	2.94	95.31	95.59	1.56	1.47
	Fine/Prefix	91.7	98.37	2.41	2.33	96.39	96.51	1.2	1.16
LARGE COEFF	Coarse/Infix	82.1	95.97	3.35	3.02	93.85	91.46	2.79	5.53
	Fine/Infix	82.5	96.44	2.86	2.56	93.71	90.77	3.43	6.67
	Coarse/Prefix	**83.5**	**96.25**	4.24	3.78	93.94	92.97	1.82	3.24
	Fine/Prefix	82	96.32	3.33	2.97	90.56	86.63	6.11	10.4
NO BACKTRACK	Coarse/Infix	75.6	94.62	2.87	3.13	93.44	86.83	3.69	10.03
	Fine/Infix	74.5	94.76	3.14	3.56	93.33	78.63	3.53	17.81
	Coarse/Prefix	**79.7**	**95.38**	7.39	6.57	89.16	83.94	3.45	9.49
	Fine/Prefix	74.7	95.23	2.37	2.41	96.44	89.16	1.19	8.43
MEDIUM DEG	Coarse/Infix	**92.8**	**98.26**	5.56	6.02	86.11	79.52	8.33	14.46
	Fine/Infix	83.4	96.12	6.63	5.56	89.76	83.33	3.61	11.11
	Coarse/Prefix	87.7	96.82	4.07	3.57	93.5	90.71	2.44	5.71
	Fine/Prefix	90.6	97.92	8.51	7.55	89.36	87.74	2.13	4.72
MEDIUM TERMS	Coarse/Infix	72.7	93.99	25.64	24.18	73.26	69.72	1.1	6.1
	Fine/Infix	75.1	95.42	21.29	20.51	75.1	69.94	3.61	9.55
	Coarse/Prefix	**76.3**	**95.78**	7.59	8.71	88.61	84.67	3.8	6.62
	Fine/Prefix	74.8	95.55	14.68	16.76	79.76	74.28	5.56	8.96

Table 4: Errors for 1 variable in the COARSE and FINE configuration for both Infix and Prefix input representation. (No curriculum, No annotation).

Config	Proof Type	Full Proof		Error Percentage					
		Full Proof Accuracy	Greedy Stepwise Accuracy	First FacStep	Total FacStep	First MulStep	Total MulStep	First SumStep	Total SumStep
SMALL COEFF	Coarse/Infix	87.9	97.01	5.79	4.9	88.43	79.72	5.79	15.38
	Fine/Infix	**88.1**	**97.19**	8.4	7.98	75.63	68.1	15.97	23.93
	Coarse/Prefix	**91.2**	**98.08**	1.14	1.03	88.64	84.54	10.23	14.43
	Fine/Prefix	90.3	97.83	8.25	6.35	80.41	73.02	11.34	20.63
MEDIUM COEFF	Coarse/Infix	88.5	97.35	4.35	3.73	83.48	76.87	12.17	19.4
	Fine/Infix	82.2	96.25	2.25	1.83	76.4	68.81	21.35	29.36
	Coarse/Prefix	**84.5**	**96.03**	3.87	3.68	88.39	81.58	7.74	14.74
	Fine/Prefix	72.4	93.6	12.68	9.95	76.09	67.74	11.23	22.31
LARGE COEFF	Coarse/Infix	80.4	95.18	6.12	4.84	82.65	75.81	11.22	19.35
	Fine/Infix	73	93.85	11.85	8.74	70.74	62.3	17.41	28.96
	Coarse/Prefix	**83.7**	**96.23**	4.29	3.61	87.12	82.99	8.59	13.4
	Fine/Prefix	78.6	95.6	5.14	4.2	81.31	74.43	13.55	21.37
NO BACKTRACK	Coarse/Infix	**72.7**	**93.13**	4.4	3.15	87.55	75.79	8.06	21.07
	Fine/Infix	72.5	94.64	3.27	2.54	85.09	73.79	11.64	23.66
	Coarse/Prefix	**63.2**	**89.87**	3.26	2.24	91.3	78.73	5.43	19.03
	Fine/Prefix	60.7	90.48	2.29	1.58	89.31	72.64	8.4	25.79
MEDIUM DEG	Coarse/Infix	**80.5**	**95.13**	6.67	5.44	81.54	71.97	11.79	22.59
	Fine/Infix	73.5	94.21	7.17	6.55	68.3	57.83	24.53	35.61
	Coarse/Prefix	**83.4**	**96.41**	4.82	4.19	81.33	75.39	13.86	20.42
	Fine/Prefix	73.6	94.57	7.58	6.38	75.38	67.48	17.05	26.14
MEDIUM TERMS	Coarse/Infix	64	92.03	25	19.05	72.5	66.5	2.5	14.45
	Fine/Infix	64	92.98	13.61	8.62	79.44	69.59	6.94	21.79
	Coarse/Prefix	**67.8**	**93.58**	10.25	7.69	87.89	80.98	1.86	11.32
	Fine/Prefix	62.9	92.74	9.16	5.97	85.71	74.37	5.12	19.65

Table 5: Errors for 2 variables in the COARSE and FINE configuration for both Infix and Prefix input representation. (No curriculum, No annotation).

Config	Proof Type	Endpoint		# Train Examples	Full Proof		Stepwise		Calibration			
		# Endpoint Examples	Endpoint Accuracy		Full Proof Accuracy	Greedy Stepwise Accuracy	Top-1 Accuracy	Beam-5 Accuracy	Sure Rate	P	R	F1
SMALL COEFF	Fine	4.3M	94.7	3.6M	82.3	97.93	86.47	87.5	81.83	100	94.64	0.97
	Coarse			5.1M	**85**	**98.31**	93.5	94.03	90.27	100	96.54	0.98
MEDIUM COEFF	Fine	7M	95.3	5.4M	78.8	97.78	93.8	94.5	90.2	99.93	96.09	0.98
	Coarse			5M	**80.1**	**97.69**	89.37	90.27	86.77	99.96	97.05	0.98
LARGE COEFF	Fine	9M	91.8	4.1M	70.1	96.59	84.8	86.63	77.77	99.83	91.55	0.96
	Coarse			4M	**73.2**	**96.66**	92.77	93.8	87.23	100	94.04	0.97
NO BACKTRACK	Fine	8.6M	83.8	3.5M	**46.5**	**92.93**	84.9	87.67	74.5	99.96	87.71	0.93
	Coarse			6.7M	**65.5**	**95.7**	67.8	69.37	63.3	99.79	93.17	0.96
MEDIUM DEG	Fine	4.9M	87.9	3.9M	59.6	95.28	94.13	95.7	86.4	100	91.78	0.96
	Coarse			4.1M	**65.1**	**95.61**	85.43	87.43	78.2	99.96	91.49	0.96
MEDIUM TERMS	Fine	8.5M	90	4.8M	**56.9**	**95.7**	92.4	93.83	85.77	99.88	92.71	0.96
	Coarse			4.2M	52.8	94.57	84	85.93	75.93	99.82	90.24	0.95

Table 6: Results for FINE and COARSE configurations for 2 Variables for annotated proofs (No curriculum).

Config	Proof Type	Full Proof		Error Percentage							
		Full Proof Accuracy	Greedy Stepwise Accuracy	First FacStep	Total FacStep	First MulStep	Total MulStep	First SumStep	Total SumStep	First MarkStep	Total MarkStep
SMALL COEFF	Fine	88.5	98.82	3.48	2.99	89.57	83.58	6.09	11.19	0.87	2.24
	Coarse	**91.9**	**99.16**	1.23	1.19	98.77	96.43	0	1.19	0	1.19
MEDIUM COEFF	Fine	78.6	97.66	18.69	15.19	74.77	74.44	3.27	6.67	3.27	3.7
	Coarse	**84.2**	**98.29**	4.43	4.65	84.81	84.88	6.33	5.81	4.43	4.65
LARGE COEFF	Fine	75.5	97.37	11.43	9.21	72.65	66.35	10.61	19.68	5.31	4.76
	Coarse	**80.3**	**97.86**	5.58	5.86	90.86	87.39	1.02	4.5	2.54	2.25
NO BACKTRACK	Fine	**68**	**96.78**	7.19	6.46	86.56	78.54	5.62	12.71	0.62	2.29
	Coarse	**59.7**	**95**	6.2	5.25	88.09	76.88	3.72	15.41	1.99	2.45
MEDIUM DEG	Fine	76	97.37	11.67	10.85	82.5	80.34	3.75	6.78	2.08	2.03
	Coarse	**78.7**	**97.38**	6.1	5.84	86.38	81.32	4.69	9.73	2.82	3.11
MEDIUM TERMS	Fine	**70.4**	**97.48**	16.89	16.27	75	69.14	3.72	8.85	4.39	5.74
	Coarse	66.2	96.34	25.44	25.28	68.05	63.48	2.37	5.81	4.14	5.43

Table 7: Errors for FINE and COARSE configurations for 1 Variable for annotated proofs (No curriculum).

Config	Proof Type	Full Proof		Error Percentage							
		Full Proof Accuracy	Greedy Stepwise Accuracy	First FacStep	Total FacStep	First MulStep	Total MulStep	First SumStep	Total SumStep	First MarkStep	Total MarkStep
SMALL COEFF	Fine	82.3	97.93	4.52	3.07	86.44	68.97	7.34	24.14	1.69	3.83
	Coarse	**85**	**98.31**	2	1.68	88.67	78.21	8	18.44	1.33	1.68
MEDIUM COEFF	Fine	78.8	97.78	8.96	6.79	80.19	68.21	9.43	22.5	1.42	2.5
	Coarse	**80.1**	**97.69**	9.05	7.79	87.94	80.33	3.02	10.66	0	1.23
LARGE COEFF	Fine	70.1	96.59	13.38	10	69.9	59.32	13.38	25.45	3.34	5.23
	Coarse	**73.2**	**96.66**	10.45	7.84	79.85	70.87	7.84	18.49	1.87	2.8
NO BACKTRACK	Fine	46.5	92.93	9.16	5.15	74.21	57.9	14.58	33.69	2.06	3.25
	Coarse	**65.5**	**95.7**	3.19	2.61	90.14	77.31	5.22	18.27	1.45	1.81
MEDIUM DEG	Fine	59.6	95.28	7.43	5.48	72.03	57.1	15.35	32.9	5.2	4.52
	Coarse	**65.1**	**95.61**	6.88	5.26	78.51	67.79	11.46	24.63	3.15	2.32
MEDIUM TERMS	Fine	**56.9**	**95.7**	21.58	13.3	67.29	57.72	8.58	23.96	2.55	5.02
	Coarse	52.8	94.57	23.94	15.6	68.22	62.77	3.6	16.67	4.24	4.96

Table 8: Errors for FINE and COARSE configurations for 2 Variable for annotated proofs (No curriculum).

a setting where polynomial simplification steps only involve symbolic addition and multiplication, without any arithmetic manipulation. For example, instead of multiplying 3 and 4 as 12, the model will output $c_1 * c_2$ given coefficients c_1 and c_2. The results for 1 Variable setting are presented in Table 9. Here, MEDIUM COEFF and MEDIUM DEGREE denote the same configuration as the case with integer coefficients. The only difference being that the limits of coefficients no longer apply. The errors made by the model for each kind of step are summarized in Table 10. We observe that the proof accuracy is about 20% less than

Config	Proof Type	Endpoint		#Train	Full Proof	
		#EE	Endpoint Acc.		Full Proof Acc.	Greedy Stepwise Acc.
MEDIUM COEFF	Coarse/Infix	5M	93.5	4.3M	78.5	94.19
	Fine/Infix			2.8M	63.5	90.64
	Coarse/Prefix	4.9M	89.77	3.7M	70.9	91.53
	Fine/Prefix			4.3M	70.9	93.2
MEDIUM DEGREE	Coarse/Infix	5.6M	88	3.7M	65.2	89.55
	Fine/Infix			6.3M	75.5	94.59
	Coarse/Prefix	6.3M	83.93	3.4M	57.6	85.98
	Fine/Prefix			6.7M	67.7	92.76

Table 9: Results for Symbolic Coeff setting. (No curriculum, No annotation).

Config	Proof Type	Endpoint		#Train	Full Proof	
		#EE	Endpoint Acc.		Full Proof Acc.	Greedy Stepwise Acc.
MEDIUM COEFF	Coarse/Infix	5M	93.5	4.3M	78.5	94.19
	Fine/Infix			2.8M	63.5	90.64
	Coarse/Prefix	4.9M	89.77	3.7M	70.9	91.53
	Fine/Prefix			4.3M	70.9	93.2
MEDIUM DEGREE	Coarse/Infix	5.6M	88	3.7M	65.2	89.55
	Fine/Infix			6.3M	75.5	94.59
	Coarse/Prefix	6.3M	83.93	3.4M	57.6	85.98
	Fine/Prefix			6.7M	67.7	92.76

Table 10: Errors made by models in Symbolic Coeff setting. (No curriculum, No annotation).

H OUT-OF-DISTRIBUTION EVALUATION

We present the results for Out-of-Distribution evaluation here. Table 11 contains results for best 2 variable models (Prefix/Coarse) tested on 1 Variable setting.
Table 12 contains results for best 1 variable models (Prefix/Coarse) tested on SMALL, MEDIUM and

LARGE coefficient setting. As expected, the SMALL and MEDIUM models perform much worse when tested on higher coefficients.

Config	Train/Test= 2 Var/1 Var		Train/Test= 1 Var/1 Var		Train/Test= 2 Var/2 Var	
	Full Proof Acc.	Greedy Stepwise Acc.	Full Proof Acc.	Greedy Stepwise Acc.	Full Proof Acc.	Greedy Stepwise Acc.
SMALL COEFF	**95.34**	**99.12**	95.3	98.97	91.2	98.08
MEDIUM COEFF	87.4	97.11	**93.6**	**98.58**	84.5	96.03
LARGE COEFF	**89.4**	**97.13**	83.5	96.25	83.7	96.23
NO BACK TRACK	**84.2**	**98.29**	79.7	95.38	63.2	89.87
MEDIUM DEG	**87.7**	**97.83**	87.7	96.82	83.4	96.41
MEDIUM TERMS	**78.5**	**96.16**	76.3	95.78	67.8	93.58

Table 11: Results for OOD Testing. NVAR = 2 COARSE/PREFIX models tested on corresponding NVAR = 1 setting (No curriculum, No annotation).

Train Config	Test Config					
	SMALL COEFF		MEDIUM COEFF		LARGE COEFF	
	Full Proof Acc.	Greedy Stepwise Acc.	Full Proof Acc.	Greedy Stepwise Acc.	Full Proof Acc.	Greedy Stepwise Acc.
SMALL COEFF	95.3	98.97	33.4	69.05	31	68.02
MEDIUM COEFF	**96.6**	**99.29**	93.6	98.58	33.6	68.96
LARGE COEFF	95.8	99.1	**94.4**	**98.64**	**83.5**	**96.25**

Table 12: Prefix/Coarse 1 Variable Models tested on various coefficient limit configurations (SMALL, MEDIUM and COARSE). (No curriculum, No annotation).

I CURRICULUM LEARNING

Learning the simplification steps should entail learning the sub-tasks, such as addition and multiplication (of numeric coefficients and symbolic variables); where multiplying variables precludes learning to add exponents of similar variables. As these sub-tasks are well-defined and dependencies among them are clear, we explore different types of curriculums based on the Mastering-Rate-based (MR) curriculum learning algorithm proposed in Willems et al. (2020). Authors in Willems et al. (2020) define curriculum learning by 1) a *curriculum* i.e. a set of tasks $\mathcal{C} = \{c_1, \ldots, c_n\}$, where a task is set of examples of similar type with a sampling distribution, and 2) a *program* which for each training step defines the tasks to train the learner given its learning state and the curriculum. Formally, the program $d : \mathbb{N} \rightarrow \mathcal{D}^{\mathcal{C}}$, is a sequence of distributions over $\mathcal{C}$. The authors estimate the *program* function through an *attention* function which defines attention over the tasks at a time-step, and an *attention-to-distribution converter* which converts the attention to a distribution over $\mathcal{C}$. Authors observe that other algorithms (Matiisen et al., 2019; Graves et al., 2017) are special cases of the above setting with different choices for *program*.

To learn on tasks that are *learnable but not learnt yet*, authors define an *ordered curriculum* $\mathcal{O}^{\mathcal{C}}$ which is a directed graph over tasks in $\mathcal{C}$. An edge from A to B indicates that learning task A before B is preferable. For supervised learners, the learnability for each task depends on mastering rate ($\mathcal{M}_c(t)$) computed from the normalized mean accuracy for that task at time-step t. At each time-step, the MR algorithm computes attention over a task ($a_c(t)$) from mastering rates of its ancestors and successors. During training to sample batches, a hyperparameter N_b for the curriculum determines the number of batches to be considered at a step, before re-computing the attention over tasks.

Using the *program d*, we first sample $N_b * b$ examples from tasks in $\mathcal{C}$. The model is then trained on randomly sampled N_b minibatches are sampled updating the mastering rates.

For polynomial simplification for 1 variable, we define the following tasks ADD, MUL2, MUL3, SCOEFF and MIXED. For ADD, only one factor per product is allowed, so there is no multiplication. For MUL2 and MUL3 only 1 product is allowed with maximum two factors and three factors respectively. SCOEFF points to the SMALL COEFF configuration and MIXED is the final variable size configuration of the target variable configuration. We define the following curriculums:
• C: {(ADD, MUL3), (MUL3, MIXED), (ADD, MIXED)}.
• C2: {(ADD, MUL2), (MUL2, MUL3), (MUL3, MIXED), (ADD, MIXED)}.
• C4: {(ADD, MUL2), (MUL2, MUL3), (MUL3, SCOEFF), (ADD, SCOEFF) (SCOEFF, MIXED)}.
For all our experiments, we use the MR algorithm with gAmax Linreg A2D converter functions described in Willems et al. (2020). Model parameters and the training configurations remains the same as before[1]. We show the results in Table 13 for COARSE configuration. As coefficient size grows from SMALL, MEDIUM, LARGE to NO BACKTRACK - the improvements in full proof accuracy steadily increase from 1% to 10.84%. For NO BACKTRACK, the improvement in top-1 accuracy is by 20% from a no curriculum setting. However, we observe for MEDIUM TERMS, there is a drop in accuracy for all curriculums and input representations. It is possible that, more carefully designed curriculums may improve the results. There is no conceivable pattern observed for infix or prefix representations. However, compared to learning without curriculum, the improvement observed for infix representation is larger than prefix.

				Full Proof		Step-wise		Calibration			
		Curriculum	#Train	Full Proof Accuracy	Stepwise Accuracy	Top-1 Acc	Beam-5 Acc	Sure Rate	P	R	F1
Small Coeff	Infix	C	2.8M	94.38	98.76	94.84	96.68	89.36	100	94.22	0.97
		C2	2M	**95.98**	**99.0**	91.64	93.24	86.16	99.9	93.98	0.97
	Prefix	C	2.02M	94.26	98.65	77.76	80.46	70.62	99.94	90.77	0.95
		C2	2.29M	94.6	98.56	93.44	95.28	88.02	99.89	94.09	0.97
Medium Coeff	Infix	C2	3.9M	**95.44**	**99.02**	94.86	96.44	91.18	100	96.12	0.98
		C4	2M	93.86	98.59	88.22	90.24	84.68	99.91	95.90	0.98
	Prefix	C2	3.7M	94.78	98.82	91.98	93.66	88.08	99.93	95.69	0.98
		C4	4.4M	94.8	98.87	85.3	87.82	80.62	99.98	94.49	0.97
Large Coeff	Infix	C2	6.9M	91.26	97.92	96.4	98.06	90.24	99.89	93.51	0.97
		C4	7.6M	91.62	98.16	91.54	93.3	87.38	99.84	95.3	0.98
	Prefix	C2	6.5M	92.2	98.31	85.38	87.78	81.42	99.95	95.32	0.98
		C4	6.97M	**92.46**	**98.42**	91.3	93.34	87.54	100.0	95.88	0.98
No Backtrack	Infix	C2	4.8M	86.44	97.27	93.68	95.46	88.72	99.98	94.68	0.97
		C4	5.1M	85.96	97.21	94.64	96.1	89.5	100	94.57	0.97
	Prefix	C2	7M	86.16	97.30	82.24	84.44	77.46	99.95	94.14	0.97
		C4	5.5M	**86.48**	**97.45**	92.6	94.3	87.78	99.95	94.75	0.97
Medium Degree	Infix	C2	3.5M	87.12	97.01	84.16	87.44	78.46	99.95	93.18	0.96
		C4	3.4M	94.12	98.65	90.62	81.984	86.66	99.93	95.56	0.98
	Prefix	C2	5.35M	**94.28**	**98.71**	80.8	82.84	75.76	100	93.51	0.97
		C4	3.5M	92.38	98.30	83.7	85.48	78.94	99.92	94.24	0.97
Medium Terms	Infix	C2	4.4M	59.54	75.76	65.6	69.56	60.84	95.36	88.45	0.92
		C4	3.8M	56.94	76.72	69.84	73.44	60.76	97.5	84.82	0.91
	Prefix	C2	2.8M	41.84	51.24	40.62	45.36	36.9	92.57	84.10	0.88
		C4	3.37M	49.02	65.41	58.56	64.64	45.44	96.83	75.14	0.85

Table 13: Curriculum Learning results for 1 variable for the COARSE configuration for both Infix and prefix representations.

REFERENCES

Alex Graves, Marc G Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *International Conference on Machine Learning*, pp. 1311–1320, 2017.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning. *IEEE transactions on neural networks and learning systems*, 2019.

[1]We use N_b as 10. For other default parameters in CL, please check `github.com/lcswillems/automatic-curriculum`.

Bartosz Piotrowski, Josef Urban, Chad E. Brown, and Cezary Kaliszyk. Can neural networks learn symbolic rewriting?, 2019.

Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 37–42, 2019.

Lucas Willems, Salem Lahlou, and Yoshua Bengio. Mastering rate based curriculum learning, 2020.